COMMENTARY

Mechanistic Approaches and the Development of Alternative Toxicity Test Methods

Michael Balls

European Centre for the Validation of Alternative Methods, JRC Environment Institute, Ispra, Italy

A mechanism can be defined as an explanation of an observed phenomenon that explains the processes underlying the phenomenon in terms of events at lower levels of organization. A prerequisite for new, more mechanistic, approaches, which would use *in vitro* systems rather than conventional animal analogy models, is a strengthening of the underlying scientific basis of toxicity testing. This will require greater recognition of the differences between fidelity and discrimination models and between analogy and correlation models. The development of high-fidelity, high-discrimination tests with a sound mechanistic basis will also require greater appreciation of the interdependence of all the components of test systems and the development of new alternative (i.e., nonanimal) testing strategies that can provide the specific knowledge needed for making relevant and reliable predictions about the potential effects of chemicals and products in human beings. The optimal use of this new knowledge will require fundamental changes to current practices in risk assessment. — *Environ Health Perspect* 106(Suppl 2):453–457 (1998). *http://ehpnet1.niehs.nih.gov/docs/1998/Suppl-2/453-457balls/abstract.html*

Key words: alternative method, *in vitro* test, hazard prediction, mechanistic test, risk assessment, toxic mechanism

Introduction

There is currently much discussion about the need for mechanistic tests, although it is not always clear what this term means. For example, mechanistic tests could mean tests involving biologic systems with a mechanistic basis that is understood, or tests that are able to identify those effects mechanistically related to the *in vivo* effects to be predicted.

In fact, calls for more mechanistic approaches include a wide spectrum of thoughts, from "We must try to make toxicity testing more scientific" or "We need a specific test for the interaction with this type of receptor on this type of cell" to "We need a test for identifying chemicals that inhibit the activity of this particular enzyme."

Significant improvements to the scientific basis of toxicity tests and to strategies for the application of the information they provide are needed before we can hope to develop new mechanistic tests (indeed, new tests of any kind) of any significance.

Art or Science?

Some toxicologists [e.g., Dayan (*1*)] argue that toxicity tests should not be seen as scientific experiments in the usual sense, and because of the generality and inevitable lack of rigor of such tests, they should be considered metascience rather than true science.

Roberfroid (*2*) is less willing to accept this contrast between what he calls the science of toxicology and the art of toxicity testing. He questions the acceptability of the current reliance of risk assessment on a poor understanding of mechanisms of toxicity, on questionable theoretical models (e.g., the two-step hypothesis of carcinogenicity), or on controversial and unsubstantiated assumptions (e.g., the ability of a linear dose–effect relationship to support the extrapolation of extremely high-dose effects to low-dose effects). Roberfroid argues that new scientifically sound, mechanism-based tests are needed that take into account the most relevant molecular and cellular events that play a role in the toxicity of chemicals, and that such tests will inevitably involve batteries of replacement alternative (i.e., nonanimal) test methods.

Meanwhile, a good example of the difference between the true science of toxicology and the metascience of toxicity testing has recently been published. Ashby and his colleagues (*3*) wanted to establish that acetochlor, an herbicide of commercial importance, does not pose a genotoxic or carcinogenic hazard to humans despite the detection of some genotoxic effects and benign tumors in rats that were given the chemical at the maximum tolerated dose. An enormous but unspecified number of animals, as well as considerable human and other resources, were used in true scientific toxicology studies made necessary by embarrassing results provided by metascientific tests. The conclusion was that the effects in the rats were not relevant to humans; that is the heart of the problem. Laboratory animals will always be imperfect models, and the relevance of the data provided by routine regulatory tests can almost never be known, as studies such as those conducted by Ashby et al. (*3*) can rarely be afforded. That is why large safety factors are often used in risk assessment in an attempt to allow for inaccuracies due to species differences, high-dose to low-dose extrapolation, etc. (*4*).

Progress will also be limited unless toxicologists are prepared to address the consequences of irrational high-dose regimes. The following comment by Ashby et al. (*3*) about their own work provides a warning: "At these elevated dose levels, associated tissue changes were encountered that compromised interpretation and extrapolation of the toxicities observed to lower dose levels—in particular, to those levels of exposure likely to be encountered by humans."

Chamberlain (*5*) has put forward the case for the development of new procedures for risk assessment based on replacement alternative test data, which will mean making significant changes to the regulatory process itself. This will require a more rational and more critical approach to hazard prediction and risk assessment than is fashionable at present. If new methods provide us with better data, we must optimize the uses to which we put them. This in turn will require better understanding of the nature of models and of the strengths and limitations of the kinds of information they can provide.

This paper was prepared as background for the 13th Meeting of the Scientific Group on Methodologies for the Safety Evaluation of Chemicals (SGOMSEC): Alternative Testing Methodologies held 26–31 January 1997 in Ispra, Italy. Manuscript received at *EHP* 9 May 1997; accepted 22 August 1997.

Address correspondence to Dr. M. Balls, European Centre for the Validation of Alternative Methods (ECVAM), JRC Environment Institute, Via Enrico Fermi, 21020 Ispra (VA), Italy. Telephone: 39 332 785 996. Fax: 39 332 785 336. E-mail: michael.balls@jrc.it

Abbreviations used: QRA, quantitative risk assessment; SAR, structure–activity relationship.

Fidelity, Discrimination, Analogy, Mechanism, and Correlation

Fidelity is the accuracy with which a model reproduces the overall properties of what is being modeled, whereas discrimination is the accuracy with which a model reproduces a particular property or properties of what is being modeled.

No model can offer 100% fidelity or 100% discrimination, but the best models will have the highest possible fidelity in combination with the highest possible discrimination. In general, a low-fidelity/high-discrimination model is more likely to be useful than a high-fidelity/low-discrimination model.

Russell and Burch (*6*) warned of the high-fidelity fallacy, to expose the weakness of the following kind of argument: "Humans are mammals, so other mammals, such as monkeys, dogs or rats, are better models of humans than are other organisms, such as birds, fish, insects or bacteria." The problem is that high fidelity does not mean high discrimination. For example, certain chemicals induce peroxisome proliferation in rats, which is linked with the induction of tumors. Therefore, according to high-fidelity reasoning, rat peroxisome proliferators may be human carcinogens because of the overall similarity of rats and humans. However, this is unlikely because the chemicals that induce peroxisome proliferation in rats do not appear to do so in humans. In this instance the rat is a low-discrimination model despite its overall relatively high fidelity.

It is possible that chemicals that cause peroxisome formation in rats might interact with a specific receptor in humans and initiate tumor formation. If so, the relationship between this chemical–receptor interaction in humans and peroxisome formation in rats would have to be satisfactorily established before the rat could be accepted as a suitable model for identifying chemicals that act via such a mechanism.

The assumed relevance of animal tests is based on the general high fidelity of animal models, i.e., on analogy (where similarity in a particular circumstance is inferred from agreement or similarity in an acceptable number of other features in the systems being compared) rather than on mechanism (where similarity is based on an adequate knowledge of the mechanistic basis of the phenomenon under consideration and its operation in the systems being compared).

In any case, similarity does not mean identity, so judgment in the interpretation of the meaning of data will always be necessary, whatever the model may be.

Correlative approaches are based solely on statistical relationships between those phenomena that cannot be explained on a mechanistic basis. They are unlikely to lead to correlative tests that will receive widespread acceptance, even if they would likely be more useful than current animal analogy tests lacking a sound mechanistic basis.

Qualitative and quantitative structure-activity relationship (SAR) models represent correlative approaches, but mechanistic SAR approaches are possible, e.g., when interactions with specific receptors can be predicted from structure and the consequences of such interactions are understood.

Greater recognition of these truths would lead to a more rational view of the value of animal procedures and improve the prospects for nonanimal tests. Animal models will always be limited by inescapable species differences, and mechanistic studies in animals are so prohibitively expensive that only a small number of compounds could ever be tested in this way. Replacement alternative methods offer the possibility of greater relevance based on better science, e.g., through mechanistic tests on human material.

High-Fidelity and Mechanistic Tests

There can also be confusion over whether a high-fidelity test is a mechanistic test. For example, the use of whole rat embryos *in vitro* is an example of a high-fidelity model, as the cultured embryos are similar to rat embryos *in utero*. Thus, when whole-embryo cultures are used to screen chemicals for teratogenicity according to a number of specified, relevant end points, we have a high-fidelity test. We do not have a sufficient understanding of the cellular or molecular basis of teratogenicity for this to be a mechanistic test. Whether it is acceptable as a high-discrimination test will depend on the outcome of a formal validation study such as the one being managed by the Zentralstelle zur Erfassung und Berwetung von Ersatz- und Ergänzungsmethoden zum Tierversuch under contract to the European Centre for the Validation of Alternative Methods (*7*).

A mechanism has been defined as "an explanation of an observed phenomenon, which explains the processes underlying the phenomenon in terms of events at lower levels of organisation" (*8*). For example:

- Paracetamol (acetaminophen) overdosage can lead to death in humans; the mechanism is death as a result of hepatic failure.
- Hepatic failure as a result of paracetamol overdosage in turn results from hepatotoxicity.
- The presumed mechanism of hepatotoxicity is metabolism of paracetamol to *N*-acetyl-*p*-benzoquinoneimine.
- The presumed mechanism of hepatotoxicity of *N*-acetyl-*p*-benzoquinoneimine is covalent binding to macromolecules.
- Covalent binding to macromolecules presumably results in impairment of the functions of those molecules, leading to cell toxicity, hepatic failure, and death of the patient.

A mechanistic test is based on a system at an acceptable level of organization and a relevant end point based on a sufficient understanding of the cellular and/or molecular basis of the effect under consideration. An example is a test based on a critical or pivotal stage in the development of an effect, such as interaction with a defined receptor.

Mechanistic tests are the tests most likely to be high-discrimination tests, but the fidelity of the system must also be borne in mind. For example, the *Salmonella typhimurium* test is a relatively high-discrimination test for genotoxicity, but a liver S9 fraction must be incorporated to improve its fidelity, i.e., its ability to detect metabolism-mediated genotoxicity.

In contrast, the *S. typhimurium* test is only a low-fidelity, low-discrimination test for nongenotoxic carcinogens.

The Components of Test Procedures

The purpose of a test procedure is to gain relevant and reliable information about a defined set of circumstances as a basis for making decisions about further action.

Toxicity test procedures can have the following components: a biologic system; an end point, which refers to the processes, responses, or effects assessed; an exposure regimen; an end point measurement, which refers to the techniques used to assess end points; a data analysis method and a way of expressing the result; a prediction model, which is the tool used to convert the results from a test into a prediction of toxicity *in vivo*; and a means of expressing toxic hazard (a quantitative expression of the adverse effects elicited by a chemical under defined conditions of exposure).

The expression of hazard obtained from the performance of a relevant and reliable

test can be used in risk assessment. Risk is the probability that an event will occur given a particular condition of exposure [Figure 1; (*9*)]. It is expressed as the product of the hazard and the likelihood of exposure, where exposure is estimated for a specific population.

It must be recognized that the quality of the eventual risk assessment is entirely dependent on the strength (i.e., the relevance and reliability) of each component of the test procedure. A test cannot be better than its weakest component.

New Test Development and Validation

Validation is the process whereby the reliability and relevance of a procedure are established for a particular purpose (*10*). It is a question of demonstrating the scientific integrity and practical usefulness of new methods and their application. The relevance of the test method and the prediction model must be evaluated separately, as must their combination as a test procedure.

There is now wide international agreement that formal validation should be a crucial and unavoidable step between new test development and regulatory acceptance. The concepts of prevalidation and prediction models are currently being refined to improve the quality and speed of formal validation. The stages in the evolution of new tests (*9*) are as follows:

Test development (laboratory of origin)
- Purpose of the test
- Need for the test
- Derivation of the method
- Application to appropriate chemicals
- Case for inclusion in a validation study
- Production of a protocol
- Development of a prediction model

Prevalidation (informal interlaboratory study)
- Optimization of the test protocol
- Assessment of its interlaboratory transferability
- Optimization of the prediction model

Validation (formal interlaboratory study, including a blind trial)
- Two phases: preliminary phase (training set of chemicals) and definitive phase (test set of chemicals)
- Main stages: study design; selection of tests and laboratories; selection, distribution, and testing of chemicals; data collection and analysis; and assessment of performance of test and of applicability of prediction model

Independent assessment (of study and proposals)

Progression toward regulatory acceptance

It is commonly believed that validation is the limiting step in the acceptance of new test methods. However, it is now becoming clear that, in fact, the main limiting factor is new test development.

Recently Purchase (*11*) stated:

> The whole validation process is designed to provide very specific answers to questions such as: "Will the method predict toxicity qualitatively?" or "Will the method predict toxicity quantitatively?" In that respect there is a close analogy with the development of a new chemical product: once the new compound is synthesized, the steps of development until it is commercialised can be predicted to a certain degree. The work of method validation is akin to technology development.
>
> Test development is much less predictable. It is closer to the definition of a scientific activity in that the outcome of research is unpredictable, both in terms of content and time. To develop methods to predict a particular toxic consequence with precision requires that the mechanism of action which leads to the toxic effect is understood. A good example is the Ames' *Salmonella* mutation method which has a precise mechanistic basis for its performance. Its development relied on many years of understanding of genetics of *Salmonella* and the recognition of the mechanism of mutation induction before there was a realistic chance of a satisfactory method. Thus, test development will depend not only on the attention given by individual scientists to developing methods but also to the scientific advances over a wide area of biology. Identifying new areas of biology ripe for exploitation and having insight into the mechanisms of toxicity will be the limiting steps in test development, where the really innovative step is the synthesis of the development process, although no less demanding in terms of effort and dedication, has as its main objective to ensure that its properties are suitable for its intended use. Thus the distinctive feature of product development is at the invention stage. (*11*)

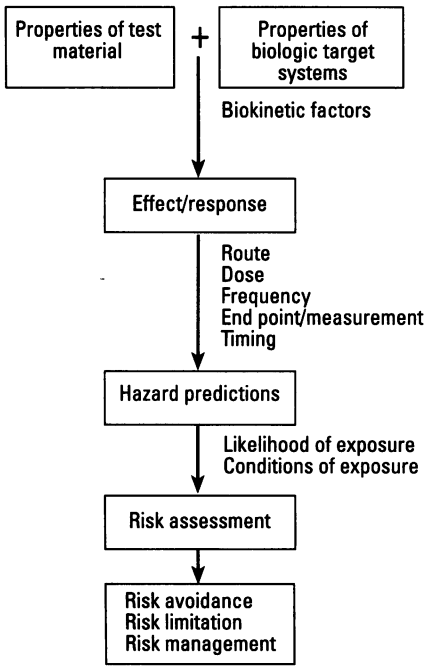

Figure 1. Testing, hazard prediction, and risk assessment. Data from Balls and Fentem (*8*), with permission of the Fund for the Replacement of Animals in Medical Experiments.

Existing Tests, New Tests, and Knowledge Needed

Much has been written about difficulties experienced in finding *in vivo* data of sufficiently high quality and relevance for use in the development and validation of *in vitro* tests. This is partly because industrial companies want to keep confidential what they consider their own property.

There is, however, another problem. If the data from an animal test are themselves of limited usefulness in terms of the purpose of the test, i.e., for predicting the likelihood of particular effects in human beings, those data must be of limited utility as a basis for evaluating the reliability of *in vitro* test data for predicting the likelihood of those effects in human beings.

It is therefore essential that there is a regular, thorough, and objective review of all animal test methods in light of the purposes for which they are used. This in turn means that there must be an objective analysis of those purposes.

If we intend to use tests to provide essential knowledge as a means of developing the safest, most effective products possible, we must first define more precisely the knowledge required. If another goal is to develop valid nonanimal test procedures for providing that knowledge, we must decide how these new test procedures should be validated. If the existing animal test is reliable and relevant in providing the necessary knowledge, data from that test can be used in the validation of potential replacement alternative methods. If not, no attempt should be made to use such data in the validation of new tests. In those circumstances, the only way forward is to

establish a convincing relationship between the information that can be provided by the nonanimal test procedure and the knowledge needed to predict likely effects in human beings.

Regulatory authorities and many toxicologists in industry are apprehensive about having to change the practices to which they have long been accustomed, however unscientific and inadequate they may be. Happily, others are more progressive, including Pioda (*12*), a Swiss regulator, who made the following statement at a conference in Zürich in 1993, on "Alternatives to Animal Testing: New ways in the Biomedical Sciences, Trends and Progress":

> It should not be forgotten that real alternatives are in essence revolutions, and revolutions cannot be incorporated into an existing structure. All our laws are based on animal experiments. Therefore, it should be acknowledged that the existing structure will have to be changed before something revolutionary can be introduced. These thoughts are particularly directed at those who are involved in the preparation and revision of laws. If this revolutionary seed of new methods is so important for the classification of chemicals, then the philosophy of the laws should be changed in accordance. Without this essential change, a breakthrough will never be achieved. (*12*)

Testing, Hazard Prediction, and Risk Assessment

Quantitative risk assessment (QRA) models have been used since the 1950s for estimating carcinogenic risks from environmental pollutants, particularly in the United States. Approximately six mathematical models are in use currently (e.g., probit, logit, and multistage models), the basis of which is the application of mathematical equations to the tumor incidence derived experimentally from long-term, typically rodent, carcinogenicity studies. A QRA equation is derived from measurements of the pharmacodynamics and kinetics of the substance in animals in combination with scaling factors. The additional use of safety factors is highly subjective and as a result, assessments based on a QRA may be in error by several orders of magnitude. Such models could possibly be validated retrospectively to some extent by using epidemiologic methods. However, this would be limited to studies of hazardous chemicals that have been present previously in the environment, where subgroups of the population with different histories of exposure could be identified.

Mathematical models may have the potential to make risk analyses more accurate, but such sophisticated techniques are not necessarily of any real value. Risk assessment is crucially dependent on the validity of the scientific assumptions made and the relevance and reliability of the toxicology data used. Nevertheless, however sophisticated the analysis may be, rats and dogs are not human beings. Therefore, in many cases mathematical models may only serve to make the assessment deceptively precise, and the conservatism of the approach currently used precludes any proper evaluation of the use of mathematical models in risk assessment. At the present time, nonspecific models based on assumptions, which may or may not be scientifically correct, are used for risk assessment. In the future we should attempt to use a quantitative, more mechanistic approach whenever possible.

This is particularly true, for example, in the case of tests for human carcinogens. The insuperable problem of species differences and the bizarre use of repeated, very high dose exposure regimes inevitably lead to a questioning of the scientific rationale and merit of the traditional rodent bioassay and of the value of any predictions based on it. There has been a revolution in concepts of carcinogenesis. This must be taken into account in new test development, which must in turn include strategies based on molecular and cell biology and on the emerging understanding of the carcinogenic process. Nowhere are theoretical and practical mechanistic approaches more necessary.

Concluding Comments

The following are among the points to be actively taken into account if we want to improve the scientific basis of toxicity tests and testing strategies:

- Agreement must be reached on what is understood and encompassed by the term mechanistic approaches.
- Approaches to toxicity testing, and to new methods in particular, must be based on a better understanding of normal physiologic and toxicologic processes.
- The development of mechanistically based tests will therefore be dependent on progress in the science of toxicology.
- Emphasis should be placed on high-discrimination tests rather than on high-fidelity models.
- The application of new approaches must be based on integrated testing strategies involving, for example, *in vitro* and quantitative SAR procedures rather than on animal models. Mechanistic approaches with animal models are too difficult and costly and are of questionable relevance where the assessment of risk to humans is the ultimate objective.
- Such new mechanistic approaches are likely to require fundamental changes to current practices in risk assessment.

REFERENCES AND NOTES

1. Dayan AD. Significance of toxicity data. In: Chemicals Testing and Animal Welfare. Solna, Sweden:National Chemicals Inspectorate, 1986;91–109.
2. Roberfroid M. Toxicology: a science and an art. Toxicol In Vitro 9:839–844 (1995).
3. Ashby J, Kier L, Wilson AGE, Green T, Lefevre PA, Tinwell H, Willis GA, Heydens WF, Clapp MJL. Evaluation of the potential carcinogenicity and genetic toxicity to humans of the herbicide acetochlor. Hum Exper Toxicol 15:702–735 (1996).
4. Balls M, Fentem JH. The use of basal cytotoxicity and target organ tests in hazard identification and risk assessment. ATLA 20:368–388 (1992).
5. Chamberlain M. The challenge of validation. In: Alternatives to Animal Testing (Lisansky SG, Macmillan R, Dupuis J, eds). Newbury:CPL Press, 1996;87–99.
6. Russell WMS, Burch RL. The Principles of Humane Experimental Technique. London:Methuen, 1959.
7. Call for laboratories to participate in the formal validation of three *in vitro* embryotoxicity tests. ATLA 25:491 (1997).
8. Frazier JM. The role of mechanistic toxicology in test method validation. Toxicol In Vitro 8:787–791 (1994).
9. Balls M, Fentem JH. Progress toward the validation of alternative tests. ATLA 25:33–43 (1997).
10. Balls M, Blaauboer B, Brusick D, Frazier J, Lamb D, Pemberton M, Reinhardt C, Roberfroid M, Rosenkranz H, Schmid B et al. Report and recommendations of the

CAAT/ERGATT workshop on the validation of toxicity testing procedures. ATLA 18:313–336 (1990).
11. Purchase IFH. How can we avoid legislation which creates a false expectation of the likely success of the search for alternatives in toxicity testing? In: Proceedings of the European Congress on the Ethics of Animal Experimentation, Brussels (in press).
12. Pioda L. The position of the authorities. In: Alternatives to Animal Testing: New Ways in the Biomedical Sciences, Trends and Progress (Reinhardt CA, ed). Weinheim:VCH Verlagsgesellschaft, 1994;173–176.